Bayesian Audio-to-Score Alignment Based on Joint Inference of Timbre, Volume, Tempo, and Note Onset Timings

AUTHOR(S):

Maezawa, Akira; Okuno, Hiroshi G.

Akira Maezawa and Hiroshi G. Okuno
Graduate School of Informatics
Department of Intelligence Science
and Technology
Kyoto University
Yoshida-Honmachi
Sakyo, Kyoto 606-8501, Japan
akira.maezawa@music.yamaha.com,
okuno@i.kyoto-u.ac.jp

Bayesian Audio-to-Score Alignment Based on Joint Inference of Timbre, Volume, Tempo, and Note Onset Timings

Abstract: This article presents an offline method for aligning an audio signal to individual instrumental parts constituting a musical score. The proposed method is based on fitting multiple hidden semi-Markov models (HSMMs) to the observed audio signal. The emission probability of each state of the HSMM is described using latent harmonic allocation (LHA), a Bayesian model of a harmonic sound mixture. Each HSMM corresponds to one musical instrument's part, and the state duration probability is conditioned on a linear dynamics system (LDS) tempo model. Variational Bayesian inference is used to jointly infer LHA, HSMM, and the LDS. We evaluate the capability of the method to align musical audio to its score, under reverberation, structural variations, and fluctuations in onset timing among different parts.

Classical music, compared with other genres of music, is unique in the extent to which a piece of music is played, many times by many people, using exactly the same set of notes. This gives rise to countless interpretations of the same piece. This variety in interpretation adds a unique way to enjoy classical music: finding the listener's favorite interpretation, or interpretations, of a given piece. This way of enjoying music, however, is often difficult for people who have just started listening to classical music, for there are an overwhelming number of recordings of any given piece of music. This point is evident, for example, when one searches for Beethoven's *Pathetique* sonata in an online store: in June 2013, there were 822 audio CDs of the sonata available at Amazon.com. The deluge of recordings makes it a challenge for a user to find "the" recording whose interpretation is matched to the listener's taste.

We seek to alleviate such a burden by inferring aspects of musical audio pertaining to interpretation. This way, interpretation-based filtering or search methods could narrow down the number of interpretations from which the user should choose. Similar motivations for comparative analysis of musical audio based on music interpretation have led to interfaces for grouping (Sapp 2007), querying (Maezawa, Goto, and Okuno 2010), and playing (Fremerey et al. 2007) various interpretations of a given piece of music. We ultimately seek to refine the quality of these interpretation-based systems by modeling various aspects of music interpretation.

To this end, we have developed a technique for audio-to-score alignment, the task of temporally aligning a musical score to an audio rendition of that score. Score alignment has been used extensively, especially in interactive applications such as automatic accompaniment (Hu, Dannenberg, and Tzanetakis 2003; Cont 2010) or automatic page-turning (Arzt, Widmer, and Dixon 2008), and in performance-analysis applications (Sapp 2007; Molina-Solana and Widmer 2010). Score-alignment methods geared toward interactivity must operate in real time and thus require online inference, whereas those aimed for performance analysis may use offline inference but must be accurate enough to uncover useful information about interpretation. We designed an accurate offline alignment method through inference of a generative model of musical audio given a score plus relevant aspects of musical interpretation. Namely, we infer (1) average volume of a note, (2) average timbre of a note, (3) pitch fluctuation of a note, (4) tempo trajectory of the musical audio, (5) fluctuation of note onsets among different parts, (6) which repeats are taken, and (7) room acoustics.

For this purpose we formulated MAHLER (Multiple Auto-regressive duration HSMMs with LHA

Computer Music Journal, 39:1, pp. 74–87, Spring 2015
doi:10.1162/COMJ_a_00284

Emission, with Reverberation inference), an offline audio-to-score alignment method based on a Bayesian inference of the aforementioned musical aspects. MAHLER consists of a dereverberation "front end" and an inference scheme of a model of the audio signal, given a score. The model contains nested hidden semi-Markov models (HSMMs), each of which corresponds to the sequence of positions that a single part plays. The HSMMs are tied together by a shared, smooth tempo-trajectory model using a linear dynamical system (LDS). Each HSMM emits a musical audio spectrogram based on latent harmonic allocation (LHA, see Yoshii and Goto 2012), a generative model of a spectral time slice described in terms of a mixture of harmonic sounds. The model jointly infers these aspects given the musical audio.

Existing Studies

Score alignment has been used extensively, mainly in two contexts: interactive applications and performance analysis. The former case must align the score in real time, while the latter may align in an offline manner in order to extract fine details of performance. In both cases, a score-alignment method must address two important design decisions: how to model a short audio fragment (e.g., a spectral time slice) given a combination of notes that are played, and how to model the sequence of such combinations of notes.

Models of spectral time slicing come in two flavors: feature-based and spectrum-based. In the former, a twelve-dimensional feature called the *chroma* vector (Fujishima 1999; Hu, Dannenberg, and Tzanetakis 2003; Orio, Lemouton, and Schwartz 2003; Joder et al. 2010; Macrae and Dixon 2010; Niedermayer and Widmer 2010) or one of its variants (Müller and Kurth 2006; Müller and Ewert 2010) is typically used. Each dimension of the chroma vector corresponds to the energy contained in a particular pitch class, e.g., C or C♯. Such a feature is robust against timbral variations because variations in harmonic structure are collapsed into a few dimensions. On the other hand, careful feature design is required, because the various ways to express the chroma vector affect the performance (Cho, Weiss, and Bello 2010).

The second approach to designing a generative model of spectral time slices assumes that the spectral time slice has been generated from a particular probability distribution (Raphael 2004; Peeling, Cemgil, and Godsill 2007; Maezawa et al. 2011). For example, one might assume that the power spectrum is generated by a zero-mean normal distribution, whose variance increases for frequency bins where we expect to observe significant signal components. This kind of approach has been used extensively in fields other than score alignment, such as multiple-f_0 estimation (Yoshii and Goto 2012) or score-informed source separation (Han and Raphael 2007; Itoyama et al. 2007; Ewert and Müller 2011, 2012). Because it uses the raw information, this kind of approach involves less ad hoc processing compared to the feature design approach. On the other hand, because of the added dimensionality, it is more difficult to attain robustness against variations in timbre or dynamics. Hence, the key to a good generative spectral model lies in designing an expressive statistical model that can cover the possible variations of timbre and dynamics.

Modeling the sequence of features also has two major approaches: dynamic time warping (DTW) and hidden Markov models (HMMs). Also, a few approaches have been proposed that use conditional random fields (Joder et al. 2010) or continuous state-space models (Duan and Pardo 2011b; Montecchio and Cont 2011; Otsuka et al. 2011). Dynamic time warping takes as input two feature sequences, along with a measure of dissimilarity between any two features. Then, DTW finds a mapping between the two sequences such that the net dissimilarity over the map is minimized, subject to various constraints (Hu, Dannenberg, and Tzanetakis 2003; Ewert, Müller, and Grosche 2009). A weakness of DTW lies in the difficulty of incorporating structural constraints; for example, it is nontrivial to incorporate structural variations such as repeats or jumps (Müller and Ewert 2008; Fremerey et al. 2009), or the notion of the expected duration within a given state.

An HMM interprets the sequence of features as an emission from a probabilistic sequence of states,

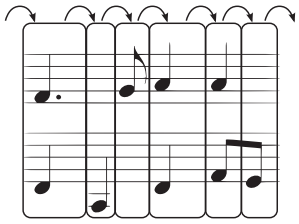
each state of which represents a unique position in the musical score. The state sequence is constrained by the state transition matrix; for example, a musical score without any repeats is represented as a transition matrix, such that the subsequent position in the score can either stay in the same place or step ahead. Moreover, various structural constraints can be imposed by manipulating the transition matrix. For example, a repeat can be trivially modeled as state transition from the end to the beginning of the repeated section. A model of state duration, such as the continuity of beat duration (Raphael 2004; Cont 2010), is achievable through hidden semi-Markov models (HSMMs).

A limitation common to these sequential models is the inability to model timing discrepancies between different parts. In other words, as long as each state corresponds to a single location in the musical score, and as long as each observation can correspond to only one state, it is impossible to convey the notion that different parts are in different states at a given time. This makes it impossible to model, say, how one part of a duo plays ahead or behind another in a rubato passage. A method to deal with this problem as a postprocessing method has been proposed by Devaney and Ellis (2009). For our problem, however, a postprocessing method such as this cannot be used, because the mutual dependency of timbre and alignment mandates a joint estimation.

Formulation of MAHLER

Our goal is to estimate the tempo, timing fluctuation between parts, timbre, dynamics, pitch, sequence of states, and reverberation. We separate this estimation into two stages. By "part," we mean an arbitrary combination of voices whose instrumentation is usually different from other parts and for which some asynchrony relative to other parts is generally expected. For example, in a piece for violin and piano, a single part may be assigned to the violin and another to the piano part. Alternately, a single part may be assigned to the violin and single part might be assigned to, say, each of the left hand and the right hand in the piano part, if we are interested

Figure 1. Model of the musical score as a state sequence.

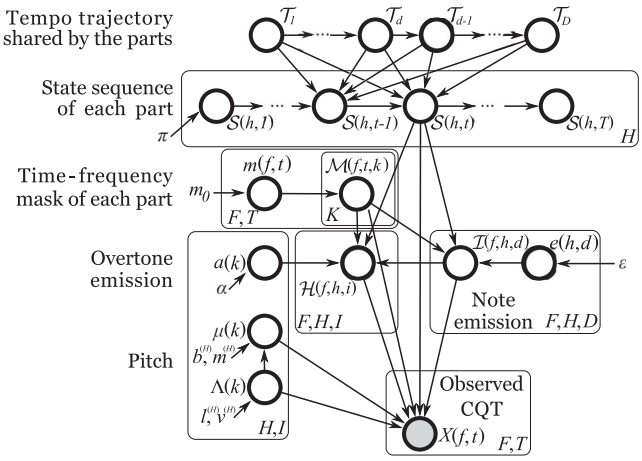


in the coordination of the left and the right hand. To achieve our goal we first use a dereverberation technique to estimate the room acoustics (late reverberation) and the "dry" audio signal that excited the room acoustics. The recovered dry signal is more representative of the musical score than is the reverberant audio, because a musical score only describes the execution of the musical piece and not the acoustics. Thus, a dereverberation front end is expected to make the alignment method robust to room acoustics. To this end, we model the observed audio in terms of a generative model of an audio spectrogram that incorporates a nonparametric Bayesian model of reverberation as a nonnegative convolution (Maezawa et al. 2014). This processing is used to jointly estimate the late reverberation and the constant Q transform (CQT) spectrogram of the dereverberated audio, $X(f, t)$, defined over frequency f and time (audio frame index) t.

Next, we jointly estimate the remaining aspects using a Bayesian model of the CQT spectrogram. We represent the score as a partition of states, by dividing the musical score vertically by note onset or offset, as shown in Figure 1. We seek to express the notion that players concur on a tempo trajectory, but musical expression and physical limitations create slight asynchrony in the execution timing among players. To this end, a sequence of states is associated with each of H constituent parts, and each sequence is expressed using an HSMM. We assume that the audio starts at the beginning of the score and ends at the end of the score. Moreover, in order to express the notion of concurred tempo trajectory, these HSMMs are tied by a single smooth tempo trajectory, expressed using an LDS.

Given such a model of the score state sequence, we model the quantized version of the dereverberated CQT $X(f, t)$. The function $X(f, t)$ is interpreted

Figure 2. Graphical model of our method. Circles represent random variables and arrows indicate conditional dependencies. See text for details.



as the number of energy quanta observed at the time-frequency bin (f, t). This kind of interpretation of an audio signal as a histogram count has been successfully applied in many tasks, such as f_0 estimation (Yoshii and Goto 2012) or sound source separation (Bryan and Mysore 2013). Each quantum is then assigned the originating part, pitch, instrument, and score position (i.e., state). We call the combination of pitch and instrument an *instrument-pitch pair* (IPP). We assume that each i of I IPPs is harmonic; thus, each IPP is assumed to emit a histogram count of the acoustic frequencies that are approximately integer multiples of the notated fundamental frequency. We express this concept using LHA because (1) it is a model of harmonic mixtures and thus works well with our model, and (2) its interpretation of the power spectrum as count data dovetails with our interpretation of X.

The overall generative process is illustrated in the graphical model of Figure 2. We shall now discuss each aspect in greater detail.

Modeling the Musical Score State Sequence

The model of the state sequence is based on first generating a smooth global tempo trajectory based on an LDS and then, for each part, creating an HSMM such that the duration of each state is governed by the LDS.

Modeling the Global Tempo Trajectory

The global tempo trajectory is designed as a smooth process, such that adjacent tempi tend to remain close to each other. We model this concept using an LDS, or an auto-regressive model of order 1. Linear dynamics systems are useful because they describe the current observation in a sequence in terms of the deviation from the previous observation; if we allow the current observation to a deviate little from the previous observation, then we can model the smoothness of the tempo trajectory.

Let $\mathcal{T}_d$ be the logarithm of the number of audio frames per tatum at state d of the sequence D; this is related to the logarithm of beats per minute (bpm). (A tatum is the greatest common divisor of the notated lengths of all the notes in the score.) Similarly to the approach taken by Raphael (2004), we assume that $\mathcal{T}_d$ deviates by a small amount from $\mathcal{T}_{d-1}$, following a normal distribution centered about $\mathcal{T}_{d-1}$:

$$\mathcal{T}_d \sim \left(\mathcal{T}_{d-1}, \mathcal{L}_{d-1}\lambda_d^{(T)^{-1}}\right). \tag{1}$$

$\mathcal{L}_d$ is the integral length, in tatums, of state d (e.g., a state whose length is a quarter note, with the tatum defined as a sixteenth note, has $\mathcal{L}_d = 4$). $\lambda_d^{(T)}$ is the precision (i.e., the inverse variance) of the tempo difference: The greater the value, the less $\mathcal{T}_d$ deviates from $\mathcal{T}_{d-1}$. We assume that this parameter is generated from the gamma distribution

$$\lambda_d^{(T)} \sim \mathcal{G}\left(l_d^{(T)}, \nu_d^{(T)}\right), \tag{2}$$

where $l_d^{(T)}$ and $\nu_d^{(T)}$ govern the distribution of the precision between $\mathcal{T}_{d-1}$ and $\mathcal{T}_d$. Specifically, the ratio $l_d^{(T)}/\nu_d^{(T)}$ is the mean of $\lambda_d^{(T)}$ and the ratio $l_d^{(T)}/\nu_d^{(T)^2}$ is its variance. Thus, they may be set as to convey the default value of $\lambda_d^{(T)}$ along with its uncertainty. For example, steady tempo can be set by assigning $l_d^{(T)}$ a value much larger than $\nu_d^{(T)}$, and making $\nu_d^{(T)}$ large, so that the expected tempo difference variance is small and the variance of $\lambda_d^{(T)}$ is small, conveying the idea that we are confident that the tempo change is small. On the other hand, upon encountering a tempo marking, these values can be set in such a

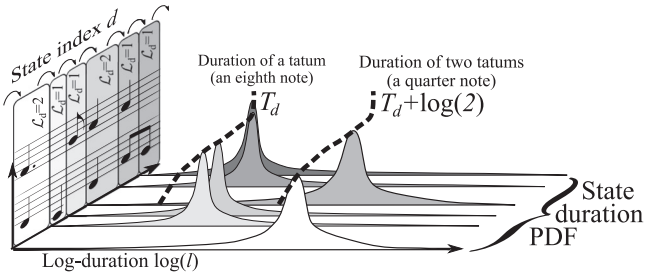


Figure 3. Model of the state duration probability distribution function (PDF).

way that $\lambda_d^{(T)}$ is distributed with a large variance, i.e., the tempo can jump by a large amount.

Modeling a Score Part as an HSMM

The musical score, as mentioned previously, is represented as a sequence of states in a manner depicted by Figure 1, and a state sequence is defined for each of the H parts. The model of the state sequence must be consistent both with the allowed transition between states and with the global tempo trajectory. The first requirement means that the transition between states must abide by the musical score, i.e., state d can transition to state $d+1$, plus repeats, cuts, or other structural markings, if any. The second requirement means that duration in each state must be explicitly modeled. To meet these requirements, the state sequence of the score part is modeled as an HSMM.

Our model is completely specified by four variables: state sequence, initial state probability density function (PDF), state transition PDF, and state duration PDF. Typically, an HSMM needs to specify the cumulative density function of the state duration PDF in order to rigorously treat the terminal condition; this information is not necessary in our model because we assume that the audio stops at the end of the score, i.e., the terminal state is constrained to be the end of the last state.

The initial state PDF π, a multinomial distribution, specifies the distribution of states at $t=1$. It is drawn from a Dirichlet distribution, i.e., $\pi|\pi_0 \sim \mathrm{Dir}(\pi_0)$. Because we assume that the music starts at the beginning of the score, and because $\langle\pi\rangle \propto \pi_0$, we set the first element of π_0 to 1, and remaining elements to a tiny positive value $\epsilon \ll 1$ (we set $\epsilon = 10^{-50}$). Other knowledge, such as optional cuts in the beginning (e.g., abridging the introductory passage of a concerto), can be coded as well.

The state transition PDF $\tau_d(d')$, a multinomial distribution, is the probability of transitioning from state d' to state d. It is also drawn from a Dirichlet distribution, i.e., $\tau(d)|\tau_0(d) \sim \mathrm{Dir}(\tau_0(d))$. The prior $\tau_0(d)$ describes the possible transitions of the musical score: It should be mostly left-to-right (i.e., the only allowed transition is from state d to $d+1$), except for few occasions where musical structure mandates otherwise (e.g., repeats). We parse the musical score to determine the appropriate $\tau_0(d)$: hyperparameters for allowed transitions are set to 1 (noninformative), and everything else is set to ϵ. Note that transition to the same state is not allowed, i.e., $\tau_{0d}(d) = \epsilon$.

The state duration PDF governs the expected number of audio frames per tatum, given the notated duration of each state. It should be consistent with the global tempo trajectory $\mathcal{T}_d$, in that the expected duration is centered about $\exp \mathcal{T}_d$, as illustrated in Figure 3. To this end, we model the log-duration at state d, $\log l$, as a normal distribution centered about the expected duration according to the tempo model. Recall that $\mathcal{L}_d$ is the number of tatums in state d and $\mathcal{T}_d$ is the log-duration of a tatum at state d as generated by the LDS. The duration PDF is then modeled as a normal distribution centered about $\mathcal{T}_d + \log \mathcal{L}_d$, with a variance of σ_T^2, $\log l \sim \mathcal{N}(\mathcal{T}_d + \log \mathcal{L}_d, \sigma_T^2)$. This kind of model allows each part to fluctuate from the global tempo trajectory, where a small value of σ_T strengthens the effect of the global tempo.

The state sequence is described as a sequence, over time (audio frame index) t, of two variables: the state d and countdown timer value l. First, at $t=1$, the HSMM chooses the initial state according to the initial state PDF, and chooses the initial countdown timer value, according to the state duration PDF at the initial state. Then, for each time $t = [1 \cdots T]$, if the current countdown timer value is not 1, the timer is decremented by 1. If the timer value is 1, the HSMM chooses the next state according to the state transition PDF associated with the current state, and chooses the countdown timer value by drawing

from the duration PDF of the next state. Let $\mathcal{S}_{l,d}(h,t)$ be a 1-of-(L, D) binary variable that indicates that the state of part h at time t is state d, with l frames remaining in d. By "1-of-K binary variable" we mean a K-dimensional unit vector, where the dimension corresponding to the active state has a binary value of one, and the other dimensions are set to zero. It is crucial to note that while the 1-of-K vector is binary, its expectation is a continuous variable. Bear in mind that this assumes that the sequence stays at a state for at most L frames. Then, $\mathcal{S}$ is described as follows:

$$p(\mathcal{S}(h,:)|\mathcal{T},\pi,\tau) = \prod_{l,d}^{L,D} \pi^{\mathcal{S}_{l,d}(h,1)} \prod_{t>1,l<L,d}^{T,L,D} 1^{\mathcal{S}_{l+1,d}(h,t-1)\mathcal{S}_{l,d}(h,t)} \times \prod_{t>1,l,d,l'\neq l+1}^{T,L,D,L} 0^{\mathcal{S}_{l',d}(h,t-1)\mathcal{S}_{l,d}(h,t)} \times \prod_{t>1,l,d,d'\neq d}^{T,L,D,D} (\tau_d(d')\mathcal{N}(\log l|\mathcal{T}_d+\log\mathcal{L}_d,\sigma_T^2))^{\mathcal{S}_{1,d'}(h,t-1)\mathcal{S}_{l,d}(h,t)}. \quad (3)$$

The last two terms of the first line expresses the countdown timer: The left term indicates that countdown timer must deterministically decrement until 1, and the right term makes any other transition illegal. The second line is activated only when the countdown timer value is 1. It expresses both the state transition and state duration; the left term in the parenthesis governs the state transition, and the right term governs the countdown timer value at the next state.

Generating a Histogram Count

Once the state sequence is generated for all H parts, the CQT spectrogram X can be generated by (1) choosing, for each count of X, the score position d, the part h, the IPP i, and the harmonics index j that generates the count and (2) emitting a quantum at some time-frequency bin (f,t) based on the choice made in (1).

First, each count is associated with one of H parts that generated it. We assume that the likelihood that a count is associated with part h follows a multinomial distribution parameterized by $m_h(f,t)$. This parameter describes the relative gain of each part at (f,t). It depends on the time and frequency because the likelihood of choosing the hth part clearly depends on t through the score position and f through the IPPs constituting the score of the hth part, which governs which frequency f is likely to be observed. At the same time, the independence of m from the actual count index means that once (f,t) is given, there is no bias towards choosing any particular quanta created inside (f,t). Because we do not know, a priori, which frequency bin is likely to appear, we set an uninformative prior on $m_h(f,t)$; this is realized by posing a Dirichlet prior with unit hyperparameter $m_{0,h}(f,t)=1$, i.e., $m_h(f,t)\sim\mathrm{Dir}(m_0(f,t))$. Then, each draw from $m_h(f,t)$ is assigned to the 1-of-H latent variable, $\mathcal{M}(f,t,k)\sim\mathrm{Mult}(m_h(f,t))$. Here, k is the actual index of the quanta at (f,t), i.e., $k\in[1,X(f,t)]$.

Next, we choose which IPP is generated. The likelihood of observing an IPP is dependent on the notated notes at each state, and on the relative volume of the notated notes. This is expressed as a multinomial likelihood of observing IPP i at state d of part h, $e_i(h,d)$. Note that $e_i(h,d)$ is independent of time; in other words, we assume that the relative volume of IPPs within a state is stationary. We assume that $e(h,d)$ is drawn from a Dirichlet distribution, $e(h,d)\sim\mathrm{Dir}(e_0(h,d))$. The prior information $e_0(h,d)$ should convey two aspects. First, the relative volume of notated IPPs should be uninformative, unless the dynamics are known through notation such as forte or piano. Second, the relative volume of IPPs that are not notated must be very close to zero. To this end, we set $e_{0i}(h,d)=1$ (uninformative) for notated IPP indices i, and set $e_{0i'}(h,d)\ll 1$ for unnotated IPP indices i'. Finally, the IPP index i drawn from from e is assigned to $\mathcal{I}_i(h,f,d)$, a 1-of-I binary variable, i.e., $\mathcal{I}(h,f,d)|e(h,d)\sim\mathrm{Mult}(e(h,d))$.

Once the IPP index i is drawn, the overtone index is drawn. We assume that each IPP is associated with a unique harmonic structure with J partials, from the fundamental up to the Jth overtone.

The harmonic structure is described as a single multinomial distribution. We set a prior distribution on $a(h,i)$, the harmonic structure of the ith IPP of part h. Specifically, we let a be a draw from a Dirichlet distribution, assign the draw to the 1-of-J binary variable $\mathcal{H}_j(h,i)$, i.e., $a(h,i) \sim \mathrm{Dir}(a_0(h,i))$, and $\mathcal{H}(h,i)|a(h,i) \sim a(h,i)$. The value of $a_0(h,i)$ determines the prior knowledge of the harmonic structure. Specifically, the relative strength of the jth overtone is proportional to the jth element of a_0, and the variance of the relative strength is governed by the norm of a_0. Therefore, we can determine a_0 in advance by training it with an instrumental sound corpus, or we can set it to an uninformative prior and determine the posterior distribution given the observed data. Note that by assuming independence from t, we assume that the relative strength of the partials remains constant within an IPP.

Next, we generate the fundamental frequency. We assume that the fundamental frequency of ith IPP, $\mu_{h,i}$, is distributed according to a normal distribution centered about the notated fundamental frequency, with precision $\lambda_{h,i}$. We set a prior distribution on $\mu_{h,i}$ such that its distribution is concentrated about the fundamental frequency of IPP i of part h, and we set λ such that its expected value is large. Specifically, we let $\mu_{h,i}$ and $\lambda_{h,i}$ be a draw from a normal-gamma distribution, i.e.,

$$\mu_{h,i}, \lambda_{h,i}|m,b,l,v \sim \mathcal{NG}(m_{h,i}^{(H)}, b_{h,i}^{(H)}, l_{h,i}^{(H)}, v_{h,i}^{(H)}).$$

$$\text{Note that } \langle \mu_{h,i} \rangle = m_{h,i}^{(H)}, \text{ and } \langle \lambda_{h,i} \rangle = l_{h,i}^{(H)} / v_{h,i}^{(H)}.$$

Finally, we draw the quanta in time-frequency bin (f,t). First, given a time t, we can choose the set of IPPs to generate the count from, by referring to the state sequence and the score. Then, the part h, IPP i and harmonics j are chosen according to the previous discussion. Then, we draw a frequency bin inside f from a normal distribution centered about the hth harmonic of the ith IPP's fundamental frequency:

$$\log f|\mathcal{S},\mathcal{H},\mathcal{I},\mathcal{M},\mu,\lambda \sim \prod_{h,i,j,t,k,d,l}^{H,I,J,T,K,D,L} \mathcal{N}(\mu_{h,i} + \log j, \lambda_{h,i}^{-1})^{\mathcal{I}_i(f,h,d)\mathcal{H}_j(f,h,i)\mathcal{S}_{l,d}(h,t)\mathcal{M}_h(f,t,k)}. \quad (4)$$

Note that the exponent activates exactly one base in the possible tuples of (h,d,l,i,j). In other words, the model assumes that each count of $X(f,t)$ is generated by a harmonic peak of a single notated IPP. Therefore, this model has the capability of jointly and uniquely identifying these elements by inferring the latent variables, i.e., $\mathcal{S}$, $\mathcal{H}$, $\mathcal{M}$, and $\mathcal{I}$.

Thus, the generative process of X can be summarized as follows:

$$p(X,\mathcal{I},\mathcal{H},\mathcal{M}|\mathcal{S},m,a,e,\mu,\lambda) = \prod_{t=1,f=1,k=1,d=1,l=1,h=1,i=1,j=1}^{T,F,X(f,t),D,L,H,I,J} (m_h(f,t)e_i(h,d)a_j(h,i)) \mathcal{N}(\log f|\mu_{h,i} + \log j, \lambda_{h,i}^{-1})^{\mathcal{I}_i(f,h,d)\mathcal{H}_j(f,h,i)\mathcal{M}_h(f,t,k)\mathcal{S}_{l,d}(h,t)}. \quad (5)$$

In other words, this is a latent variable model that associates with each (f,t,k) a latent variable of tuple (h,i,j,d,l) factored into the form $\mathcal{I}_i(f,h,d)\mathcal{H}_j(f,h,i)\mathcal{M}_h(f,t,k)\mathcal{S}_{l,d}(h,t)$.

Model Inference

Having defined the model, our goal is to determine the posterior distribution $p(\mathcal{S},\mathcal{H},\mathcal{M},\mathcal{I},\mathcal{T},e,a,\mu,\lambda,m|X)$. We can then use the statistics of the posterior distribution for our needs. For example, the maximum a posteriori estimate of the state sequence $\mathcal{S}$ becomes the score alignment.

We seek to find an approximate posterior distribution using the variational Bayes (VB) method. Approximation is necessary because straightforward application of Bayes' rule to find the posterior, i.e., $p(\Theta|X) = p(X,\Theta)/\int p(X,\Theta)d\Theta$, is impractical, as the denominator is intractable. VB approximates the posterior by minimizing the Kullback-Leibler (KL) divergence from the true posterior to an approximate posterior. Specifically, we approximate $p(\mathcal{S},\mathcal{H},\mathcal{M},\mathcal{I},\mathcal{T},e,a,\mu,\lambda,m|X)$ as a factored form $q_{\mathcal{S}}(\mathcal{S})q_{\mathcal{H}}(\mathcal{H})\,q_{\mathcal{M}}(\mathcal{M})q_{\mathcal{I}}(\mathcal{I})q_{\mathcal{T}}(\mathcal{T})q_e(e)q_a(a)q_\mu(\mu)q_\lambda(\lambda)q_m(m)$. We call this factorized approximation the *variational posterior*. By factorizing the posterior in this manner, posterior inference can be performed by iteratively minimizing the KL divergence from

the true posterior to the variational posterior with regard to each factor of the variational posterior.

A presentation of the full derivation is beyond the scope of this article. Therefore, we only present one primary mathematical contribution here, the update of $q_T(T)$, an LDS whose emission is a state-duration PDF of an HSMM, and $q_S(\mathcal{S})$, the HSMM whose duration PDF is governed by the LDS. Variational inference of other aspects is omitted because they can be derived using standard VB techniques. The full derivation is available online at winnie.kuis.kyoto-u.ac.jp/members/amaezaw1.

The variational posterior of the LDS, $q_{\mathcal{T}}(\mathcal{T})$, is determined by modifying the Kalman smoothing algorithm so as to emit histograms. Minimizing the KL divergence from the posterior to the variational posterior with regard to $q_{\mathcal{T}}(\mathcal{T})$ leads to the following equation that resembles an LDS, where $\psi_l(d) \triangleq \sum_{h,t>1,d'\neq d} \langle \mathcal{S}_{0,d'}(h,t-1)\mathcal{S}_{l,d}(h,t)\rangle_{q_S}$:

$$\log q_{\mathcal{T}}(\mathcal{T}) \stackrel{c}{=} \sum_{d=1}^{D}\left[\sum_{l=1}^{L}\psi_l(d)\log\mathcal{N}\left(\log\frac{l}{\mathcal{L}_d}\middle|\mathcal{T}_d,\sigma_T^2\right) + \left\langle\log\mathcal{N}\left(\mathcal{T}_d|\mathcal{T}_{d-1},\mathcal{L}_{d-1}\lambda_d^{(T)^{-1}}\right)\right\rangle\right]. \quad (6)$$

The equation is highly similar to an LDS in that the second term expresses the continuity of $\mathcal{T}_d$, but different in that the log-emission probability of the first term consists of all possible state durations l, weighed by $\psi(d)$, which is the expected unnormalized histogram of state duration of the nested HSMMs.

Similar to the Kalman smoother, we update the $q(\mathcal{T})$ using a forward-backward algorithm. The forward algorithm is described as the following forward recursion:

$$\begin{aligned}\alpha_d^{(\mathcal{T})}(\mathcal{T}_d) &\triangleq p(\mathcal{T}_d|\psi(1:d)) = \mathcal{N}(\mathcal{T}_d|u_d,s_d)\\ &\propto \int \alpha_{d-1}^{(\mathcal{T})}(\mathcal{T}_{d-1})\mathcal{N}(\mathcal{T}_d|\mathcal{T}_{d-1},\mathcal{L}_{d-1}\lambda^{(L)}{}_d^{-1})\\ &\quad\times\prod_{l=1}^{L}\mathcal{N}\left(\log\frac{l}{\mathcal{L}_d}\middle|\mathcal{T}_d,\sigma_T^2\right)^{\psi_l(d)}d\mathcal{T}_{d-1}. \quad (7)\end{aligned}$$

Integrating $\mathcal{T}_{d-1}$ out and completing the square with respect to $\mathcal{T}_d$ gives the following for u_d and s_d, where $m_d^{-1} = \frac{1}{s_{d-1}} + \frac{\langle\lambda_d^{(T)}\rangle}{\mathcal{L}_{d-1}}$:

$$s_d^{-1} = \frac{\sum_{l=1}^{L}\psi_l(d)}{\sigma_T^2} + \frac{\langle\lambda_d^{(T)}\rangle}{\mathcal{L}_{d-1}} - m_d\left(\frac{\langle\lambda_d^{(T)}\rangle}{\mathcal{L}_{d-1}}\right)^2; \quad (8)$$

$$u_d = s_d\left(m_d\frac{\langle\lambda_d^{(T)}\rangle}{\mathcal{L}_{d-1}}\frac{u_{d-1}}{s_{d-1}} + \sum_{l=1}^{L}\frac{\psi_l(d)}{\sigma_T^2}\log\frac{l}{\mathcal{L}_d}\right). \quad (9)$$

The backward algorithm is described as the following backward recursion:

$$\begin{aligned}\beta_d^{(\mathcal{T})}(\mathcal{T}_d) &\triangleq p(\psi(d+1:T)|\mathcal{T}_d) = \mathcal{N}(\mathcal{T}_d|v_d,q_d)\\ &= \int \beta_{d+1}^{(\mathcal{T})}(\mathcal{T}_{d+1})\mathcal{N}(\mathcal{T}_{d+1}|\mathcal{T}_d,\mathcal{L}_d\lambda_{d+1}^{(T)^{-1}})\\ &\quad\times\prod_{l=1}^{L}\mathcal{N}\left(\log\frac{l}{\mathcal{L}_{d+1}}\middle|\mathcal{T}_{d+1},\sigma_T^2\right)^{\psi_l(d+1)}d\mathcal{T}_{d+1}. \quad (10)\end{aligned}$$

By completing the square, we obtain the following, where $n_d^{-1} = \frac{1}{q_{d+1}} + \frac{\langle\lambda_{d+1}^{(T)}\rangle}{\mathcal{L}_d} + \frac{\sum_{l=1}^{L}\psi_l(d+1)}{\sigma_T^2}$:

$$q_d^{-1} = \frac{\langle\lambda_{d+1}^{(T)}\rangle}{\mathcal{L}_d} - n_d\left(\frac{\langle\lambda_{d+1}^{(T)}\rangle}{\mathcal{L}_d}\right)^2; \quad (11)$$

$$v_d = n_d q_d\frac{\langle\lambda_{d+1}^{(T)}\rangle}{\mathcal{L}_d}\left(\sum_{l=1}^{L}\frac{\psi_l(d+1)}{\sigma_T^2}\log\frac{l}{\mathcal{L}_{d+1}} + \frac{v_{d+1}}{q_{d+1}}\right). \quad (12)$$

Using these, we obtain the variational posterior as follows:

$$\begin{aligned}q(\mathcal{T}_d|l_{1:T}) &= \alpha_d^{(\mathcal{T})}(\mathcal{T}_d)\,\beta_d^{(\mathcal{T})}(\mathcal{T}_d)\\ &= \mathcal{N}\left(\mathcal{T}_d|\frac{1}{q_d^{-1}+s_d^{-1}}\left(\frac{v_d}{q_d}+\frac{u_d}{s_d}\right),\frac{1}{q_d^{-1}+s_d^{-1}}\right). \quad (13)\end{aligned}$$

This distribution can be used to generate the duration PDF of the HSMM, namely:

$$\begin{aligned}\log\mathrm{Dur}(l,d) &\triangleq \left\langle -\frac{1}{2\sigma_T^2}(\log l/\mathcal{L}_d - \mathcal{T}_d)^2 - \log(2\pi\sigma_T^2)\right\rangle_{\mathcal{T}_d} \\ &= -\frac{1}{2\sigma_T^2}\left(\log\frac{l}{\mathcal{L}_d} - \frac{1}{q_d^{-1}+s_d^{-1}}\left(\frac{v_d}{q_d}+\frac{u_d}{s_d}\right)\right)^2 \\ &\quad - \frac{1}{2\sigma_T^2(q_d{}^{-1}+s_d{}^{-1})} - \log(2\pi\sigma_T^2). \end{aligned} \tag{14}$$

Minimizing the KL divergence from the posterior to variational posterior with regard to $q_{\mathcal{S}}(\mathcal{S}|h,:)$ leads to the following:

$$\begin{aligned}\log q_{\mathcal{S}}(\mathcal{S}) = &\sum_{l=1,d=1}^{L,D}\Bigg[\mathcal{S}_{l,d}(h,1)\langle\log\pi_{l,d}\rangle \\ &+\sum_{t=2,d'\neq d}^{T,D,}\mathcal{S}_{1,d'}(h,t-1)\mathcal{S}_{l,d}(h,t)(\langle\log\tau_d(d')\rangle+\log\mathrm{Dur}(l,d)) \\ &+\sum_{t=1,f=1}^{T,F}\mathcal{S}_{l,d}(h,t)\left(\sum_{k=1}^{X(f,t)}\langle\mathcal{M}_h(f,t,k)\rangle\right)\log\kappa_d(h,f)\Bigg]. \end{aligned} \tag{15}$$

where $\log\kappa_d(h,f) = \langle\sum_{i,j}\mathcal{I}_{i,h}(f,d)\mathcal{H}_j(f,h,i)\log(e_i(h,d)a_j(h,i)\mathcal{N}(\log f/j|\mu_{h,i},\lambda_{h,i}^{-1}))\rangle$. This has the same functional form as an HSMM, with $\kappa_d(h,f)$ being an unnormalized histogram that shows the spectrum that state d is likely to emit, and $\sum_{k=1}^{X(f,t)}\langle\mathcal{M}_h(f,t,k)\rangle$ is the expected CQT spectrogram of part h. Thus, like an HMM, the Baum-Welch algorithm can be used to infer the state expectation. Let $\alpha_{l,d}^{(\mathcal{S})}(h,t) \triangleq p(\mathcal{S}_{l,d}(h,t)=1|X(f,1\cdots t))$ be the forward variable and $\beta_{l,d}^{(\mathcal{S})}(h,t) \triangleq p(X(f,t+1\cdots T)|\mathcal{S}_{l,d}(h,t)=1)$ be the backward variable of the HSMM. Then, we obtain the following recurrence, where $O_d(h,t) = \prod_{f=1}^{F}\kappa_d(h,f)^{\sum_{k=1}^{X(f,t)}\langle\mathcal{M}_h(f,t,k)\rangle}$ is the pseudo-emission probability of state d at time t:

$$\begin{aligned}\alpha_{l,d}^{(\mathcal{S})}(h,t) \propto\ & O_d(h,t)\Big(\alpha_{l+1,d}^{(\mathcal{S})}(h,t-1) \\ &+\sum_{d'=1}^{D}e^{\langle\log\tau_d(d')\rangle+\log\mathrm{Dur}(l,d)}\alpha_{1,d'}^{(\mathcal{S})}(h,t-1)\Big); \end{aligned} \tag{16}$$

$$\beta_{l,d}^{(\mathcal{S})}(h,t) = \begin{cases} O_d(h,t+1)\beta_{l-1,d}^{(\mathcal{S})}(h,t+1) & l>1 \\ \displaystyle\sum_{d'=1}^{D}O_{d'}(h,t+1)e^{\langle\log\tau_{d'}(d)\rangle} & l=1 \\ \quad\times\displaystyle\sum_{l'=1}^{L}\beta_{l',d'}^{(\mathcal{S})}(h,t+1)\,e^{\log\mathrm{Dur}(d',l')} & \end{cases}. \tag{17}$$

Because we assume that the state ends at the downbeat of the terminal state, we set $\beta_{l,d}^{(\mathcal{S})}(h,T)=1$ for $d=D$ and $l=1$, and 0 everywhere else. Based on these variables, the following expectations are given:

$$\langle\mathcal{S}_{l,d}(h,t)\rangle \propto \alpha_{l,d}^{(H)}(h,t)\,\beta_{l,d}^{(H)}(h,t); \tag{18}$$

$$\begin{aligned}&\langle\mathcal{S}_{d',0}(h,t-1)\mathcal{S}_{d,d}(h,t)\rangle \\ &\propto \alpha_{1,d'}^{(H)}(h,t-1)\,O_d(h,t)e^{\langle\log\tau_{d'}(d)\rangle+\log\mathrm{Dur}(d)}\beta_{l,d}^{(H)}(h,t). \end{aligned} \tag{19}$$

This information is then used to compute the expectation in Equation (6), and to find an estimated trajectory of part h at time t, $\arg\max_{l,d}\langle\mathcal{S}_{l,d}(h,t)\rangle$.

For a piece of about four minutes, the inference program, written in Python and some inlined C++ code, takes about 20 minutes to run on a PC with Intel Core i5 processor running at 2.6 GHz, and it uses about 1 GB of memory.

Evaluation

To evaluate MAHLER, we assess its ability to (1) align the musical score to the audio, assuming that performers play synchronously, (2) detect whether or not a given repeat sign is repeated in performance, (3) detect timing discrepancies among different players that are not notated in the score, and (4) align the musical score under reverberant acoustical conditions.

In the subsequent evaluations, we evaluated the CQT at every eighth of a semitone from C1 to E8 (where A4 = 440 Hz), at a rate of 20 frames per second. Unless otherwise noted, $a_0(h,i)$ are uninformative for all h and i, and are evaluated

up to the $J = 5$th partial. Additionally, $m_{h,i}^{(H)}$ is set to the notated fundamental frequency, $l_{h,i}^{(H)} = 10^{20}$, and $v_{h,i}^{(H)} = b_{h,i}^{(H)} = (1/2)^2 10^{20}$; this sets the prior distribution of $\mu_{h,i}$ concentrated about the notated fundamental frequency with a standard deviation of a quarter tone. Also, $\langle \lambda_d^T \rangle$ is set to 10^{-9} for all d except for the very beginning and the end, which are set to 10. Thus, the tempo LDS is allowed to vary more in the very beginning and the end. Furthermore, σ_T is swept from 10 to 0.06 over the course of VB iterations. This effectively weakens the influence of the tempo LDS model in the initial iterations of the inference. These values were set manually, based on a real-world musical audio recording that was unused in the evaluation owing to the lack of reliable ground truth data. Moreover, we first ran our model with all parts collapsed into a single part (i.e., a score with $H = 1$), and we used the variational posterior of this model to initialize the proposed model with each HSMM for each part. The maximum duration to stay in a state was set to 1 sec, or $L = 20$.

Global Alignment

This experiment compares the effectiveness of our method for audio-to-score alignment to that of three other methods: (1) a reference method based on DTW that minimizes the net cosine distance between chroma vector of the audio and the score ("Chr"), (2) our method without the LDS tempo model, where note durations are assumed to be mutually independent, and set to $p(\mathcal{T}_d) = \delta(\mathcal{T}_d - 10)$ for all d ("LH"), and (3) our method with a fixed timbre/volume/part balance ("m-MAHLER"). The DTW path constraints used in Method 1 are identical to those used by Hu, Dannenberg, and Tzanetakis (2003). Method 2 is similar to that of Peeling, Cemgil, and Godsill (2007), in that the method is dependent on a fixed tempo, but the note durations between notes are independent. For Method 3, we fix the expected emission spectrum in the HSMM, that is, we fix $\kappa_d(f)$ in Equation (15) into a fixed, normalized spectrum in a manner similar to the approach used by Raphael (2004).

First, we synthesized 60 pieces from the musical scores (in standard MIDI file format, SMF) provided by the "real world computing" (RWC) classical music database (Goto 2004), using the Freepats patch (Walsh 2013) and expressive data as entered in the SMFs. Then, we computed the expected beat positions, as well as the estimated beat positions obtained from score alignment. We evaluated the percentile of the absolute error of the beat position. This method gives an accurate ground truth, and reflects well the performance in real-world signals with human players, as suggested by Müller and Ewert (2010). To test with real audio, we also evaluated our method on ten instrumental performances of Bach chorales (Duan and Pardo 2011a).

The result is shown in Table 1. The data suggests two conclusions. First, the LDS tempo model improves the accuracy: Our method performs much better than LH. Second, treating timbre and volume in a Bayesian manner is critical: m-MAHLER performs far worse than any other methods, including the reference DTW alignment.

Evaluation of Musical Structure Inference

This experiment evaluated the ability of MAHLER to detect prenotated musical structures such as repeats and cuts. Because cuts and repeats are mathematically dealt with in the same way (as nonadjacent state transitions of the HSMM), it suffices to evaluate the capability to detect repeats. For each number of polyphonic voices p from two to six, we created 100 random scores of polyphony p with a duration of 256 beats, and a repeated segment in the middle of length 2, 4, 8, 16, and 32 bars long. These files were then synthesized at 100 bpm using a piano patch. The audio files were then aligned against the musical score. We deemed the estimation to be "correct" if the alignment result had the same number of backwards skips as the number of repeats. The left-hand side of Table 2 shows the result as a function of the repeated segment's duration, averaged over polyphony. The table shows that as the polyphony increases, the detection capability decreases. This behavior mirrors

Table 1. Percentile of Absolute Error in Milliseconds

		25%	*50%*	*75%*	*90%*	*95%*
Piano solo	Chr	90	304	1,363	6422	11,736
	LH	17	48	224	891	2,040
	m-MAHLER	1,485	4,520	10,468	19,415	26,728
	MAHLER	9	21	50	126	269
Instrument plus piano accompaniment	Chr	68	182	619	2,714	9,848
	LH	14	32	86	255	473
	m-MAHLER	863	2,549	6,437	9,373	11,219
	MAHLER	8	21	45	93	163
Small ensemble	Chr	90	259	891	2,804	4,710
	LH	16	46	131	393	816
	m-MAHLER	1,927	4,296	8,827	16,260	25,178
	MAHLER	10	22	45	88	133
Orchestral	Chr	123	394	1,384	6,688	36,550
	LH	38	104	574	4,793	16,768
	m-MAHLER	3,111	10,463	21,788	34,275	44,847
	MAHLER	23	51	119	805	2,996
Bach10 dataset (real performance)	Chr	66	205	340	1297	1988
	LH	11	30	60	166	350
	m-MAHLER	159	350	667	978	1213
	MAHLER	11	29	57	165	366

Smaller error values imply greater accuracy. See main text for discussion of the different methods evaluated.

Table 2. True Positive Rate of Repeated Segment Detection

Beat length	*2*	*4*	*8*	*16*	*32*	*Number of voices*	*2*	*3*	*4*	*5*	*6*
True positive	96%	87%	83%	89%	94%	True positive	100%	100%	99%	88%	63%

True negative rate was 100%. The tables show true positive rate as a function of the repeated segment length (left) and as a function of the number of polyphonic voices (right).

the result from Table 1, where genres with increased instrumentation performed relatively poorly. The right-hand side of Table 2 shows the result as a function of polyphony, averaged over segment duration. From this, we note that the accuracy decreases for a repeat of about 8 to 12 beats. In practice, such repeats are rare, so it should not hinder the merit of incorporating prenotated nonadjacent state transitions.

Evaluation of Onset Timing Differences between Parts

In this experiment, we evaluated the capability of MAHLER to find small unnotated onset timing fluctuations among different parts. For each number of polyphonic voices p, we synthesized 1000 SMF files, each of which consecutively plays three chords with p notes, each with a duration of 1 sec and played by a piano patch. Notes were assigned in one of p parts, such that all p parts were monophonic. Additionally, the onset times were dilated by uniformly distributed noise between −250 and 250 msec. Each of the $p \times 1000$ audio files was aligned to the score using (1) a single HSMM for the entire part, and (2) p HSMMs, one defined for each part. We compared the cumulative percentage of the absolute error of the estimated onsets. The result, presented in Table 3, shows the percentage of frames whose error lies within an error margin, for different

Table 3. Cumulative Error in Polyphony

Method	p	*10 msec*	*20 msec*	*50 msec*	*100 msec*	*200 msec*	*500 msec*	*1000 msec*
Single HSMM	2	20%	46%	57%	76%	94%	99%	100%
Our method	2	25%	58%	66%	80%	94%	99%	99%
Single HSMM	3	14%	37%	48%	70%	93%	100%	100%
Our method	3	18%	46%	56%	74%	92%	99%	100%
Single HSMM	4	10%	29%	41%	66%	92%	99%	100%
Our method	4	16%	39%	50%	70%	91%	99%	100%
Single HSMM	5	9%	27%	39%	66%	92%	99%	100%
Our method	5	14%	37%	48%	69%	90%	99%	100%

Summary of percentage of absolute error of estimated onsets with regard to number of polyphonic voices *p*.

polyphony. It shows that our method almost always has a higher percentage of frames whose error lies within a given bound, meaning that our method is more capable of detecting asynchrony between parts than when assuming that every part is completely synchronized.

Evaluation of Reverberation Robustness

This experiment evaluates the effectiveness of the dereverberation front end for aligning reverberant audio signals. To this end, three kinds of audio signals were generated using duets of a single instrument and piano from the RWC database, totaling ten pieces. First, we prepared synthesized audio without reverberation. Next, another set of audio signals was created by convolving the dry audio with a concert hall impulse response (Merimaa, Peltonen, and Lokki 2005), whose reverberation time is approximately 3 sec (s1_r1_o). Finally, another set of audio signals was created by dereverberating the reverberant audio. We aligned the musical score to the three kinds of audio signals and then evaluated the alignment error percentile. The result, shown in Table 4, shows that reverberation significantly deteriorates the alignment performance. It also shows that quality of alignment for the dereverberated audio is almost identical to the dry case. This suggests that dereverberation is critical for attaining robustness against hall acoustics.

Table 4. Percentile of Absolute Error in Milliseconds

	25%	*50%*	*75%*	*90%*	*95%*
Ideal (clean signal)	12	26	66	199	420
Without dereverberation	15	39	113	279	572
With dereverberation	12	28	68	170	310

Conclusion

This article presented MAHLER, a two-stage Bayesian audio-to-score alignment method involving dereverberation, followed by joint estimation of timbre, dynamics, tempo, pitch, score alignment, and onset timing differences among parts.

Future work will include estimation of sustain pedals and other aspects that increase the discrepancy between the musical score and the aligned audio. Also, the timing model might be improved by, for example, treating parts as coupled Markov models, instead of a nested model conditioned on a global LDS. Application to musicological studies and information retrieval of classical music are also interesting future directions.

Acknowledgments

Akira Maezawa is currently also affiliated with Yamaha Corporation, R&D Division. Hiroshi G. Okuno is currently affiliated with Waseda

University. This research was partially supported by Grant-in-Aid for Scientific Research (KAKENHI) 24220006.

References

Arzt, A., G. Widmer, and S. Dixon. 2008. "Automatic Page Turning for Musicians via Real-Time Machine Listening." In *Proceedings of the European Conference on Artificial Intelligence*, pp. 241–245.

Bryan, N. J., and G. J. Mysore. 2013. "An Efficient Posterior Regularized Latent Variable Model for Interactive Source Separation." In *Proceedings of the International Conference on Machine Learning*, pp. 208–216.

Cho, T., R. J. Weiss, and J. P. Bello. 2010. "Exploring Common Variations in State of the Art Chord Recognition Systems." In *Proceedings of the Sound and Music Computing Conference*, pp. 1–8.

Cont, A. 2010. "A Coupled Duration-Focused Architecture for Real-Time Music-to-Score Alignment." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(6):974–987.

Devaney, J., and D. P. W. Ellis. 2009. "Handling Asynchrony in Audio–Score Alignment." In *Proceedings of the International Computer Music Conference*, pp. 29–32.

Duan, Z., and B. Pardo. 2011a. "Soundprism: An Online System for Score-Informed Source Separation of Music Audio." *IEEE Journal of Selected Topics in Signal Processing* 5(6):1205–1215.

Duan, Z., and B. Pardo. 2011b. "A State Space Model for Online Polyphonic Audio-Score Alignment." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 197–200.

Ewert, S., and M. Müller. 2011. "Score-Informed Voice Separation for Piano Recordings." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 245–250.

Ewert, S., and M. Müller. 2012. "Using Score-Informed Constraints for NMF-Based Source Separation." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 129–132.

Ewert, S., M. Müller, and P. Grosche. 2009. "High Resolution Audio Synchronization Using Chroma Onset Features." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 1869–1872.

Fremerey, C., et al. 2007. "A Demonstration of the SyncPlayer System." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 131–132.

Fremerey, C., et al. 2009. "Sheet Music-Audio Identification." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 645–650.

Fujishima, T. 1999. "Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music." In *Proceedings of the International Computer Music Conference*, pp. 464–467.

Goto, M. 2004. "Development of the RWC Music Database." In *Proceedings of the International Congress on Acoustics*, vol. I, pp. 553–556.

Han, Y., and C. Raphael. 2007. "Desoloing Monaural Audio using Mixture Models." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 145–148.

Hu, N., R. B. Dannenberg, and G. Tzanetakis. 2003. "Polyphonic Audio Matching and Alignment for Music Retrieval." In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 185–188.

Itoyama, K., et al. 2007. "Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 57–60.

Joder, C., et al. 2010. "An Improved Hierarchical Approach for Music-to-Symbolic Score Alignment." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 39–44.

Macrae, R., and S. Dixon. 2010. "Accurate Real-Time Windowed Time Warping." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 423–428.

Maezawa, A., M. Goto, and H. G. Okuno. 2010. "Query-by-Conducting: An Interface to Retrieve Classical-Music Interpretations by Real-Time Tempo Input." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 477–482.

Maezawa, A., et al. 2011. "Polyphonic Audio-to-Score Alignment Based on Bayesian Latent Harmonic Allocation Hidden Markov Model." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 185–188.

Maezawa, A., et al. 2014. "Nonparametric Bayesian Dereverberation of Power Spectrograms Based on Infinite-Order Autoregressive Processes." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(12):1918–1930.

Merimaa, J., T. Peltonen, and T. Lokki. 2005. "Concert Hall Impulse Responses, Pori, Finland: Reference." Available online at www.acoustics.hut.fi/projects/poririrs. Accessed 27 January 2013.

Molina-Solana, M., and G. Widmer. 2010. "Evidence for Pianist-Specific Rubato style in Chopin Nocturnes." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 225–230.

Montecchio, N., and A. Cont. 2011. "A Unified Approach to Real Time Audio-to-Score And Audio-to-Audio Alignment Using Sequential Montecarlo Inference Techniques." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 193–196.

Müller, M., and S. Ewert. 2008. "Joint Structure Analysis With Applications to Music Annotation and Synchronization." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 389–394.

Müller, M., and S. Ewert. 2010. "Towards Timbre-Invariant Audio Features for Harmony-Based Music." *IEEE Transactions on Audio, Speech, and Language Processing* 18(3):649–662.

Müller, M., and F. Kurth. 2006. "Enhancing Similarity Matrices for Music Audio Analysis." In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 9–12.

Niedermayer, B., and G. Widmer. 2010. "A Multi-Pass Algorithm for Accurate Audio-to-Score Alignment." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 417–422.

Orio, N., S. Lemouton, and D. Schwartz. 2003. "Score Following: State of the Art and New Developments." In *Proceedings of the of the International Conference on New Interfaces for Music Expression*, pp. 36–41.

Otsuka, T., et al. 2011. "Real-Time Audio-to-Score Alignment Using Particle Filter for Coplayer Music Robots." *EURASIP Journal on Advances in Signal Processing* 2011(1):13. Available online at asp.eurasipjournals.com/content/2011/1/384651. Accessed July 2014.

Peeling, P., A. Cemgil, and S. Godsill. 2007. "A Probabilistic Framework for Matching Music Representations." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 267–272.

Raphael, C. 2004. "A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 387–394.

Sapp, C. S. 2007. "Comparative Analysis of Multiple Musical Performances." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 2–5.

Walsh, E. A. 2013. "Freepats Project." Available online at freepats.zenvoid.org. Accessed 1 August 2013.

Yoshii, K., and M. Goto. 2012. "A Nonparametric Bayesian Multipitch Analyzer Based on Infinite Latent Harmonic Allocation." *IEEE Transactions on Audio, Speech, and Language Processing* 20(3):717–730.